SVR - Technisch rapport 2006 / 3

Een statistische analyse van een toenemende of dalende ongelijkheid in participatie

Van kruistabellen naar oddsratio's en van oddsratio's naar een logistische regressie (en terug)

Jan Pickery

Studiedienst van de Vlaamse Regering

Vlaamse overheid

Een statistische analyse van een toenemende of dalende ongelijkheid in participatie

Van kruistabellen naar oddsratio's en van oddsratio's naar een logistische regressie (en terug)

Jan Pickery

Samenstelling
Diensten voor het Algemeen
Regeringsbeleid
Studiedienst van de Vlaamse Regering

Jan Pickery

Verantwoordelijke uitgever
Josée Lemaître
Administrateur-generaal
Boudewijnlaan 30
1000 Brussel

Lay-out cover
Diensten voor het Algemeen
Regeringsbeleid
Communicatie
Patricia Van Dichel

Druk
Departement Bestuurszaken
Reprografie

Depotnummer
D/2006/3241/299

Bestellingen
Caroline Temmerman
Tel. 02 553 57 84
http://publicaties.vlaanderen.be

I N H O U D S T A F E L

1. Inleiding

De Studiedienst van de Vlaamse Regering (SVR) wil bijdragen aan een kwaliteitsverhoging van de statistiekproductie binnen de Vlaamse overheid. Onze brochure "Kwaliteitszorg in het statistische productieproces" bevat heel wat aanbevelingen in verband met het verzamelen, verwerken en documenteren van statistische gegevens (APS, 2003). Verder willen wij ook concretere handleidingen aanbieden over het juiste gebruik van statistische technieken. Voorliggend document is hiervan het eerste voorbeeld.

In deze tekst proberen we duidelijk te maken, hoe je kan detecteren of een bepaalde ongelijke participatie toe- of afneemt. Voorbeelden van zo'n ongelijke participatie zijn de arbeidsdeelname van mannen en vrouwen of de cultuurparticipatie van hoger en lager opgeleiden. Voor beide voorbeelden is het gekend dat de participatie ongelijk is. Vaak zijn er ook cijfers over die ongelijkheid voor verschillende jaren. Maar of en hoe die cijfers geïnterpreteerd kunnen worden als een toename of een daling van ongelijkheid is niet altijd eenduidig.

De bedoeling van deze tekst is om een correcte weergave en interpretatie van dergelijke cijfers zo eenvoudig mogelijk en praktisch toepasbaar voor te stellen. In een bijlage wordt ook getoond hoe de verschillende bewerkingen en testen uitgevoerd kunnen worden in Excel en, indien Excel ontoereikend is, in SPSS.

Heel de tekst door volgen we hetzelfde voorbeeld. Dat voorbeeld heeft betrekking op de adaptatie van internet en de zogenaamde digitale kloof. Meer concreet gaan we na hoe het gebruik van internet verschilt tussen mensen met betaald werk en mensen zonder betaald werk én of dat verschil de laatste jaren toegenomen of eerder afgenomen is.

We gebruiken hiervoor data van de SCV-survey (survey Sociaal-Culturele Verschuivingen, voorheen APS-survey). In die survey werd in 2001, 2003 en 2005 het internetgebruik bevraagd[1]. De surveys bevatten voor ons bruikbare gegevens voor respectievelijk 1446, 1429 en 1514 respondenten.

Een eerste tabel maakt duidelijk dat in die periode het internetgebruik sterk is toegenomen, van een goede 34% internetgebruikers in 2001 tot meer dan 58% in 2005 (zie tabel 1).

Tabel 1 Evolutie van het internetgebruik in Vlaanderen

	Internetgebruik			
	ja		neen	
	%	aantal	%	aantal
2001	34,2	495	65,8	951
2003	46,5	665	53,5	765
2005	58,5	885	41,5	629

Bron: SCV-survey

[1] De vraagstelling was niet exact dezelfde in de drie surveys. In 2001 en 2003 werd gevraagd naar het regelmatig gebruik ("minstens éénmaal per maand"). In 2005 werd gevraagd naar het gebruik op zich en werden een aantal antwoordcategorieën aangeboden ("nooit", "meer dan een jaar geleden", "tussen de drie maanden en een jaar geleden", "gedurende de laatste drie maanden"). In een vervolgvraag werd dan nog eens de frequentie bevraagd ("minder dan één keer per maand", "minstens één keer per maand",...). Om de antwoorden vergelijkbaar te maken, combineren we voor 2005 beide vragen en kijken we naar de mensen die gedurende de laatste drie maanden minstens éénmaal per maand internet gebruikten.

Op deze tabel kunnen we een chikwadraattest uitvoeren, die sterke significantie zal tonen ($p<0.0001$). Strikt genomen vertelt die test ons enkel dat de kenmerken "internetgebruik" en "jaar" niet onafhankelijk zijn. In de praktijk kan je op basis van deze test wel zeggen dat het aantal internetgebruikers significant toegenomen is tussen 2001 en 2005 (wat niet zo'n verrassing is natuurlijk).

Interessanter voor onze vraagstelling (de evolutie van de ongelijkheid in internetgebruik tussen werkenden en niet-werkenden) is het verschil in internetgebruik tussen mensen met betaald werk en mensen zonder betaald werk. We bekijken dat verschil voor de drie surveyjaren in tabel 2. De tabel bevat meer informatie dan strikt noodzakelijk, maar die overtollige informatie wordt achteraf wel gebruikt in de berekeningen.

Tabel 2 Verschil in internetgebruik tussen mensen met en mensen zonder betaald werk

	Internetgebruik								
	JA				NEEN				
	betaald werk				betaald werk				
	ja		neen		ja		neen		
	%	aantal	%	aantal	%	aantal	%	aantal	
2001	47,4	367	19,0	127	52,6	408	81,0	543	***
2003	62,4	465	29,1	199	37,6	280	70,9	485	***
2005	76,4	639	36,3	246	23,6	197	63,7	432	***

*** p < 0.001, test voor verschil tussen mensen met en mensen zonder betaald werk

Bron: SCV-survey

Voor de drie jaren merken we dat er een groot verschil in internetgebruik is tussen mensen met betaald werk en mensen zonder betaald werk. In 2001 bvb. gebruikte meer dan 47% van de mensen met betaald werk internet, terwijl dit bij de mensen zonder betaald werk slechts 19% was. Chikwadraattesten maken verder duidelijk dat dat verschil telkens sterk significant is.

Zowel bij mensen met, als bij mensen zonder betaald werk stijgt het aantal internetgebruikers. Maar een inschatting van de evolutie van dat verschil is minder eenvoudig. Stijgt of daalt de ongelijkheid? We kunnen op verschillende manieren naar tabel 2 kijken. Die worden in de volgende paragrafen één voor één toegelicht.

2. Het absolute verschil tussen percentages

De meest eenvoudige manier om de evolutie van de ongelijkheid te bekijken is gewoon het absolute verschil in percentages te berekenen en dat absolute verschil te vergelijken voor de 3 jaren. Zo kunnen we narekenen dat het aandeel internetgebruikers in 2001 bij de mensen met betaald werk 28,4 procentpunten hoger lag dan bij mensen zonder betaald werk (47,4 - 19,0). In 2003 bedroeg dat verschil 33,3 procentpunten, terwijl het in 2005 opgelopen was tot 40,1 procentpunten[2]. Op basis van deze meting kunnen we concluderen dat de ongelijkheid toeneemt. Maar is deze inschatting wel volledig? Het absolute verschil houdt immers helemaal geen rekening met de ***relatieve voorsprong***, die de mensen met betaald werk hebben. *Hoeveel groter is het aandeel internetgebruikers* bij de mensen met betaald werk? De tweede methode houdt wel rekening met deze relatieve voorsprong.

[2] Bemerk dat we telkens spreken over procentpunten en niet over procenten.

3. Verhouding van percentages

De tweede methode bekijkt voor de verschillende jaren hoeveel meer (of minder) gebruikers er zijn in een bepaalde groep in verhouding tot de andere groep. Zo kunnen we zien dat er in 2001 bijna 2,5 keer zoveel internetgebruikers zijn bij de mensen met betaald werk als bij de mensen zonder betaald werk (47,4/19,0 = 2,49). In 2003 is die verhouding gedaald tot 2,14 en in 2005 tot 2,10. Volgens deze methode lijkt de relatieve voorsprong van de mensen met betaald werk dus af te nemen.

De berekening van de relatieve voorsprong op deze manier is echter problematisch. Ten eerste is de berekening te sterk afhankelijk van de oorspronkelijke percentages bij de eerste momentopname. Omdat van de mensen met betaald werk al een kleine helft internet gebruikt in 2001 is het vrijwel onmogelijk dat die mensen met betaald werk dezelfde relatieve groei laten optekenen als de mensen zonder betaald werk, waarvan er slechts 1/5 internet gebruiken in 2001. Ten tweede krijgen we andere resultaten als we zouden vertrekken van de relatieve achterstand van de mensen zonder betaald werk. In 2001 zijn er bij die mensen zonder betaald werk ongeveer 1,5 keer zoveel niet-gebruikers van internet als bij de mensen met betaald werk (81,0/52,6). In 2003 vergrootte die relatieve achterstand tot 1,9 en in 2005 tot meer dan 2,7.

Volgens deze methode zou dus enerzijds de relatieve voorsprong van de mensen met betaald werk dalen, maar zou anderzijds de relatieve achterstand van mensen zonder betaald werk stijgen. Deze paradox maakt duidelijk dat de berekeningswijze niet voldoet. Door te werken met odds en oddsratio's kan hieraan verholpen worden.

4. Odds en oddsratio's

Odds zijn kansverhoudingen. In deze context is het meer in het bijzonder de verhouding van de kans op internetgebruik tot de kans op niet-gebruik. Omdat gebruik "ja/neen" een dichotome variabele is, is de kans op niet-gebruik – uitgedrukt in percentages – natuurlijk gelijk aan 100 - de kans op gebruik:

$$\text{odds} = \frac{\%\,\text{internetgebruik}}{\%\,\text{niet-gebruik internet}} = \frac{\%\,\text{internetgebruik}}{100 \quad \%\,\text{internetgebruik}}$$

In 2001 bedroeg deze odds voor mensen met betaald werk 0,900 (47,4/52,6); voor mensen zonder betaald werk was dat 0,234 (19,0/81,0).

Alhoewel een odds een te weinig gebruikte grootheid is, kan zij toch eenvoudig en intuïtief geïnterpreteerd worden. Die interpretatie luidt: voor iedere persoon met betaald werk die geen internet gebruikt, zijn er 0,90 personen met betaald werk die wel internet gebruiken. En: voor iedere persoon zonder betaald werk die geen internet gebruikt, zijn er 0,23 personen zonder betaald werk die wel internet gebruiken. Deze interpretatie maakt ook duidelijk dat er bij een odds van 1 evenveel participanten als niet-participanten zijn.

De odds kunnen ook omgekeerd berekend worden: 52,6/47,4 = 1,11. De interpretatie wordt dan eenvoudigweg: voor iedere persoon met betaald werk die in 2001 internet gebruikte, waren er 1,11 personen die geen internet gebruikten. Bemerk ook dat 1,11 gelijk is aan 1/0,90.

Odds kunnen en mogen met elkaar vergeleken worden. Als we teruggrijpen naar de eerste berekening kunnen we stellen dat voor mensen met betaald werk de odds gelijk waren aan 3,85 keer dezelfde odds bij de mensen zonder betaald werk. Deze verhouding van twee odds wordt ook een oddsratio genoemd. Ook hier is 1 het referentiecijfer. Bij een oddsratio van 1 is de participatie van beide groepen gelijk – wat daarom niet wil zeggen dat de

participatie in beide groepen gelijk is aan 50%. Hoe meer de oddsratio verschilt van 1, hoe groter de ongelijkheid[3].

Hieronder geven we de berekening van de oddsratio's nog eens stap voor stap weer.

$$\text{oddsratio} = \frac{\text{odds (mensen met betaald werk)}}{\text{odds (mensen zonder betaald werk)}}$$

$$= \frac{\text{kans op internetge bruik (met betw)} \,/\, \text{kans op niet gebruik (met betw)}}{\text{kans op internetge bruik (zonder betw)} \,/\, \text{kans op niet gebruik (zonder betw)}}$$

$$= \frac{47{,}4 \,/\, 52{,}6}{19{,}0 \,/\, 81{,}0} = 3{,}85$$

In 2003 was dezelfde oddsratio gelijk aan 4,05 en in 2005 was dat 5,70. (De geïnteresseerde lezer kan dit eventueel zelf narekenen). Deze oddsratio's kunnen ondubbelzinnig en zonder fout als een maat van ongelijkheid geïnterpreteerd worden. De conclusie luidt dus dat de ongelijkheid in internetgebruik tussen mensen met en mensen zonder betaald werk toegenomen is van 2001 tot 2005.

Een inhoudelijke interpretatie die nauw aansluit bij de berekende maat, kan als volgt luiden: de verhouding van het aantal mensen dat internet gebruikt op het aantal mensen dat geen internet gebruikt, was in 2001 bij mensen met betaald werk gelijk aan 3,85 keer dezelfde verhouding bij mensen zonder betaald werk. In 2005 was diezelfde factor opgelopen tot 5,70. De ongelijkheid neemt bijgevolg toe!

Op die toename van de ongelijkheid kan je overigens ook een cijfer[4] plakken. Je kan stellen dat de ongelijkheid is toegenomen met een factor 1,48 (= 5,70/3,85).

Eén van de voordelen van deze oddsratio is dat hij volledig symmetrisch is. Als we zouden vertrekken van de niet-gebruikers bij de mensen zonder betaald werk, bekomen we exact dezelfde resultaten. De interpretatie luidt dan: de verhouding van het aantal mensen dat geen internet gebruikt op het aantal mensen dat wel internet gebruikt, was in 2001 bij mensen zonder betaald werk gelijk aan 3,85 keer dezelfde verhouding bij mensen met betaald werk. Deze factor loopt op tot 5,70 in 2005, wat duidt op een toegenomen ongelijkheid.

Je kan de odds en oddsratio's ook in de andere richting berekenen. Bij de berekening van de odds kan je het niet-gebruik in de teller zetten en toch de mensen met betaald werk opnemen in de teller van de formule voor de oddsratio. Gegeven de vraagstelling is dat minder interessant, maar de conclusies zijn wel equivalent. In 2001 bekomen we dan bvb. een oddsratio die gelijk is aan 0,26. In 2005 is dezelfde oddsratio gelijk aan 0,18, wat tot de volgende interpretatie leidt. In 2001 was de verhouding van de kans op niet-gebruik versus de kans op gebruik bij mensen met betaald werk gelijk aan 0,26 keer dezelfde verhouding bij mensen zonder betaald werk. Deze factor daalt tot 0,18 in 2005. Ook hier is de conclusie dat de ongelijkheid toeneemt. Bij volledige gelijkheid neemt de oddsratio immers de waarde aan van 1 en hoe verder verwijderd van 1 hoe groter de ongelijkheid.

[3] Bemerk wel dat de mogelijke waarden voor een odds of een oddsratio niet symmetrisch verdeeld zijn rond 1. Zij kunnen waarden aannemen van 0 tot +∞.

[4] Voor de volledigheid: dit wordt soms een hogere-orde-oddsratio genoemd. Het is de verhouding van twee odssratio's.

Deze conclusie is dus helemaal equivalent aan die van hierboven. Bemerk ook dat 0,26 gelijk is aan $\frac{1}{3,85}$ en 0,18 gelijk is aan $\frac{1}{5,70}$.

Om dit deel over odds en oddsratio's af te sluiten, tonen we nog dat een berekening met aantallen juist dezelfde resultaten oplevert als een berekening met percentages. We tonen dit voor de eerste oddsratio die we hierboven berekend hebben.

$$\text{oddsratio} = \frac{367/408}{127/543} = \frac{0.8995}{0.2339} = 3,85$$

We hebben nu aangetoond dat oddsratio's een goede maat zijn om de evolutie van een ongelijke participatie te bekijken. We kunnen ook testen of deze oddsratio's significant zijn. Dat vergt echter een omweg[5] die buiten het bereik van deze tekst valt. In de volgende paragraaf tonen we wel hoe de resultaten van een logistische regressie volledig gelijkwaardig zijn aan deze berekening van oddsratio's. Bovendien geeft deze logistische regressie, uitgevoerd met gespecialiseerde software, ook automatisch significantietesten.

5. Logistische regressie

Binaire logistische regressie[6] is een statistische techniek die geschikt is om dichotome afhankelijke variabelen te analyseren. De afhankelijke variabele van het model heeft dus twee categorieën, zoals bvb. "gebruikt internet" en "gebruikt internet niet". De onafhankelijke variabelen van het model kunnen zowel numeriek als categorisch zijn. Het is hier niet de plaats om een uitgebreide inleiding te geven in de techniek van logistische regressie. Zo'n inleiding kan eventueel elders gevonden worden[7]. Wij beperken ons ertoe om te tonen welke informatie uit de output van een logistische regressie gehaald kan worden die relevant is voor onze vraagstelling hier.

In een eerste stap tonen we een logistische regressie voor de data van 2001, met internetgebruik als afhankelijke variabele en het al dan niet hebben van betaald werk als onafhankelijke variabele. Als een logistische regressie maar één categorische onafhankelijke variabele bevat, geeft die eigenlijk dezelfde informatie als een kruistabel met chikwadraattest, zij het in andere vorm.

Voor deze logistische regressie hebben we zowel betaald werk als internetgebruik omgezet in dummies, variabelen die uitsluitend de waarden 0 en 1 kunnen aannemen. De waarde 0 staat voor respectievelijk géén internetgebruik en géén betaald werk, de waarde 1 voor het tegenovergestelde. Bij de onafhankelijke variabelen wordt de categorie die waarde 0 gekregen heeft, ook de referentiecategorie genoemd.

Als we die 0/1-variabelen in een logistische regressie stoppen, krijgen we het resultaat dat weergegeven wordt in tabel 3.

[5] Die omweg loopt via de asymptotische standaardfout van de log van de oddsratio.

[6] De afhankelijke variabele van een logistische regressie kan ook meerdere categorieën tellen. Dan spreekt men van multinomiale logistische regressie.

[7] Een korte inleiding kan gevonden worden bij Pampel (2000). Een meer uitvoerige en diepgaande tekst is deze van Hosmer en Lemeshow (2000).

Tabel 3 Resultaten van de logistische regressie van de kans op internetgebruik voor 2001

	B	S.E.	Wald	df	Sig.	Exp(B)
betw	1,347	,122	121,846	1	,000	3,846
Constante	-1,453	,099	217,277	1	,000	,234

Bron: SCV-survey

In eerste instantie is het vooral de laatste kolom die ons interesseert. Die kolom toont de geëxponentieerde parameters van de logistische regressievergelijking[8]. Deze geëxponentieerde parameters vormen eigenlijk een odds en een oddsratio. De exponent van de parameter die hoort bij de constante (of het intercept van de vergelijking) is gelijk aan 0,234 en is de odds die geldt voor de referentiecategorie van de onafhankelijke variabele, mensen zonder betaald werk. We hadden eerder ook al berekend dat die odds gelijk was aan 0,23.

Odds kan je overigens ook omzetten in een proportie of percentage, met de volgende formule:

$$p = \frac{\text{odds}}{1+\text{odds}} = \frac{0,234}{1+0,234} = 0,190$$, wat dus overeenkomt met 19,0 % uit de tabel.

De geëxponentieerde parameter voor betaald werk is 3,846. Dat is de oddsratio (die we ook al hierboven gevonden hadden). Die oddsratio maakt duidelijk hoeveel groter of kleiner de odds is voor de andere categorie (in dit geval dus voor de mensen met betaald werk).

De odds voor mensen met betaald werk is dus gelijk aan $0,234 \times 3,846 = 0,900$ wat overeenkomt met een percentage gelijk aan $\frac{0,900}{1+0,900} = 0,474$ of 47,4%.

De informatie van de logistische regressie is dus vergelijkbaar met die van de kruistabel. In deze fase levert deze logistische regressie weinig of geen toegevoegde waarde. We zien wel dat de oddsratio significant is (de voorlaatste kolom), maar op basis van de chikwadraattest wisten we ook al dat het verschil tussen mensen met en mensen zonder betaald werk significant was in 2001.

Om de evolutie van de ongelijkheid in te schatten zullen we een logistische regressie uitvoeren op de samengevoegde data. We voegen de datasets van 2001, 2003 en 2005 samen in één dataset, en nemen daarin ook een jaarvariabele op. De dataset die we zo krijgen bevat uiteindelijk 4.389 personen. Daarvan gebruikt er zo'n 47% internet en heeft iets minder dan 54% betaald werk (zie tabel 4 en 5). Tabel 6 toont dat onze dataset telkens ongeveer 1/3 personen bevat van 2001, 2003 en 2005.

Tabel 4 Internetgebruik voor de drie jaren samen

	Frequentie	Percentage
,00	2345	53,4
1,00	2043	46,6
Total	4388	100,0

Bron: SCV-survey

[8] De parameter zelf vind je in de eerste kolom (B). De geëxponentiëerde parameter is exp(B) of e^B.

Tabel 5 Al dan niet hebben van betaald werk voor de drie jaren samen

	Frequentie	Percentage
,00	2032	46,3
1,00	2356	53,7
Total	4388	100,0

Bron: SCV-survey

Tabel 6 Jaar van bevraging voor de samengevoegde dataset

	Frequentie	Percentage
2001,00	1445	32,9
2003,00	1429	32,6
2005,00	1514	34,5
Total	4388	100,0

Bron: SCV-survey

Op zich zijn deze tabellen natuurlijk wat "verdoezelend". Het internetgebruik vertoont eigenlijk te grote verschillen over de drie jaren om het zomaar in één tabel weer te geven. Maar die informatie over de jaarverschillen zit natuurlijk in de dataset en kan er eenvoudig uitgehaald worden met een kruistabel of een logistische regressie.

We willen nu in één analyse de oddsratio voor 2001, 2003 en 2005 berekenen en tegelijkertijd zien of die oddsratio's significant van elkaar verschillen. Dat kan in een logistische regressie. De afhankelijke variabele is (opnieuw) internetgebruik. Onafhankelijke variabelen zijn betaald werk en jaar. Jaar moet dan ook gehercodeerd worden (bvb. naar 2 dummies).

Als we die variabelen echter gewoon zouden opnemen in het model, zouden we een soort van gemiddelde effecten berekenen: bijvoorbeeld het gemiddelde verschil over de drie jaren heen tussen mensen zonder en mensen met betaald werk of het gemiddelde verschil tussen 2003 en 2001 (los van het al dan niet hebben van betaald werk). Dat is echter niet onze vraagstelling. Wij willen het verschil tussen werkenden en de niet-werkenden in 2001 én datzelfde verschil in 2003 én in 2005. Dat kunnen we bekomen als we interactie-effecten modelleren. Interactie-effecten worden berekend door de producttermen van de verschillende variabelen op te nemen in de analyse. Deze hele procedure zullen we nu stapsgewijze verduidelijken.

Eerst maken we dummy-variabelen (0/1-variabelen) aan voor jaar. Als we een categorische variabele omzetten in dummies, behouden we altijd één dummy minder dan het oorspronkelijke aantal categorieën in de data. Wij hebben gegevens voor 2001, 2003 en 2005, we zullen dus twee dummies overhouden. De referentiecategorie kunnen we arbitrair kiezen. hier lijkt het logisch om het eerste jaar (2001) te nemen. We maken dus een dummy jaar03 en jaar05 aan. Personen die in 2001 bevraagd werden, krijgen twee keer de waarde 0 op beide dummies. Personen die in 2003 bevraagd werden, krijgen waarde 1 op jaar03 en waarde 0 op jaar05 en voor personen die in 2005 in de survey zaten, geldt net het omgekeerde. Tabel 7 toont het codeerschema.

Tabel 7 Dummycodering van het jaar van bevraging

		Jaar van bevraging		
		2001	2003	2005
nieuwe dummy	jaar03	0	1	0
	jaar05	0	0	1

Tabel 8 en tabel 9 geven de frequentieverdeling van beide dummy-variabelen. Een vergelijking met tabel 6 maakt nog eens duidelijk hoe de codering toegepast is.

Tabel 8 Frequentietabel van de dummy voor het jaar 2003 ("jaar03")

	Frequentie	Percentage
,00	2959	67,4
1,00	1429	32,6
Total	4388	100,0

Tabel 9 Frequentietabel van de dummy voor het jaar 2005 ("jaar05")

	Frequentie	Percentage
,00	2874	65,5
1,00	1514	34,5
Total	4388	100,0

We kunnen nu een logistische regressie laten lopen met internetgebruik als afhankelijke variabele en betaald werk en jaar03 en jaar05 als onafhankelijke variabelen. De resultaten van die logistische regressie vind je in tabel 10.

Tabel 10 Resultaten van de logistische regressie van de kans op internetgebruik voor de drie jaren samen (met betaald werk en jaar als onafhankelijke variabelen)

	B	S.E.	Wald	df	Sig.	Exp(B)
betw	1,504	,067	500,456	1	,000	4,500
jaar03	,607	,082	54,751	1	,000	1,835
jaar05	1,110	,082	184,103	1	,000	3,034
Constante	-1,557	,074	443,446	1	,000	,211

Bron: SCV-survey

Als we de resultaten in tabel 10 bekijken, vinden we geen enkel cijfer terug dat overeenstemt met onze berekening van odds en oddsratio's in paragraaf 4. De reden hiervoor is al aangehaald. De effecten van "betw", "jaar03" en "jaar05" zijn netto-effecten, effecten die controleren voor de andere effecten opgenomen in het model. Zo geeft "betw" het effect van het hebben van betaald werk op het al dan niet gebruiken van het internet, als de effecten van "jaar03" en "jaar05" onder controle worden gehouden. Met een beetje goede wil kan je dit ook interpreteren als een gemiddeld effect van betaald werk op het internetgebruik over de drie jaren heen. Omdat we juist geïnteresseerd zijn in een verschillend effect van betaald werk op het internetgebruik voor de drie jaren, schiet dit model tekort. Een specifiek effect voor elk jaar kunnen we bekomen door interactie-effecten op te nemen in het model.

Interactietermen kunnen in regressies opgenomen worden als producttermen. Deze producttermen zijn gewoon vermenigvuldigingen van de dummy betaald werk met de twee dummies voor jaar (jaar03 en jaar05). Als je twee 0/1-variabelen vermenigvuldigt, is het resultaat natuurlijk terug een 0/1-variabele. Dat blijkt ook uit tabel 11 en tabel 12 die de frequentieverdelingen van de producttermen weergeven.

Tabel 11 Frequentietabel van de productterm van betaald werk en jaar03 ("betw03")

	Frequentie	Percentage
,00	3643	83,0
1,00	745	17,0
Total	4388	100,0

Tabel 12 Frequentietabel van de productterm van betaald werk en jaar05 ("betw05")

	Frequentie	Percentage
,00	3552	80,9
1,00	836	19,1
Total	4388	100,0

Deze tabellen geven eigenlijk gewoon het aantal mensen dat in het respectievelijke jaar bevraagd werd én toen betaald werk had. In tabel 2 kan je gaan kijken hoeveel mensen in 2003 betaald werk hadden. De groep internetgebruikers en niet-gebruikers daarvan tel je op en je komt aan 745. Die 745 personen hebben waarde 1 gekregen op de dummy "betw03" en alle andere personen in de samengevoegde dataset kregen waarde 0. Voor 2005 kan dezelfde berekening gemaakt worden.

In onze volgende regressie kunnen we nu als onafhankelijke variabele zowel de effecten van betw, jaar03 en jaar05 opnemen als de producttermen (betw03 en betw05). Die logistische regressie geeft de resultaten van tabel 13.

Tabel 13 Resultaten van de logistische regressie van de kans op internetgebruik voor de drie jaren samen (met betaald werk en jaar en de producttermen van die variabelen als onafhankelijke variabelen)

	B	S.E.	Wald	df	Sig.	Exp(B)
betw	1,347	,122	121,846	1	,000	3,846
jaar03	,562	,130	18,803	1	,000	1,754
jaar05	,890	,127	49,194	1	,000	2,435
betw03	,051	,166	,094	1	,759	1,052
betw05	,393	,167	5,527	1	,019	1,481
Constant	-1,453	,099	217,277	1	,000	,234

Bron: SCV-survey

Deze resultaten komen nu wel overeen met onze vroegere berekeningen. We kijken terug naar de laatste kolom. De exponent van het intercept van de vergelijking is de odds die geldt voor de referentiecategorie van de onafhankelijke variabelen, mensen zonder betaald werk voor 2001. We hadden inderdaad ook al berekend dat die odds gelijk was aan 0,234. Het eerste cijfer in die kolom 3,85 is de oddsratio voor 2001. De interpretatie, die we ook hierboven al gegeven hebben, luidt dus: in 2001 waren er voor iedere persoon zonder betaald werk die geen internet gebruikte 0,23 personen die wel internet gebruikten. Bij de

mensen met betaald werk, was dezelfde verhouding 3,85 keer hoger. Zij gebruiken dus relatief veel vaker het internet.

De oddsratio's voor jaar03 en jaar05 vertellen ons ***hoeveel groter de odds*** is voor mensen zonder betaald werk in respectievelijk 2003 en 2005. Dat is voor onze vraagstelling terug wat minder interessant.

Wel relevant zijn de oddsratio's voor betw03 en betw05. Zij tonen ***hoeveel groter de oddsratio*** is in respectievelijk 2003 en 2005. De verhouding die we hierboven berekend hebben voor 2001 (3,85), was in 2003 dus nog 1,05 keer groter en in 2005 1,48 keer groter. Zo komen we bij een oddsratio van 4,05 in 2003 ($3{,}846 \times 1{,}052$) en 5,70 in 2005 ($3{,}846 \times 1{,}481$). Deze cijfers die we ook al berekend hadden, tonen dus dat de ongelijkheid is toegenomen: een klein beetje van 2001 tot 2003, maar duidelijk van 2001 tot 2005 (met een factor vrijwel gelijk aan 1,5).

De logistische regressie geeft ons bovendien informatie over de significantie van die toenames. Die vinden we in de voorlaatste kolom. De parameter van betw03 blijkt niet significant op niveau α = 0,05, die van betw05 wel. Het verschil in internetgebruik tussen mensen zonder betaald werk en mensen met betaald werk is significant groter in 2005 dan datzelfde verschil in 2001. In 2003 was het daarentegen ***niet*** significant groter of kleiner dan in 2001. Zo kunnen we dus onze conclusie verder verfijnen: de ongelijkheid in het gebruik van het internet volgens het al dan niet hebben van betaald werk, is niet significant toegenomen (noch afgenomen) van 2001 tot 2003. Maar in 2005 is die ongelijkheid wel significant groter dan in 2001. We kunnen dus zeker niet stellen dat mensen zonder betaald werk aan een inhaaloperatie bezig zijn. Ondanks het feit dat ook bij hun het internetgebruik stijgt (zie bvb. tabel 2), is het tegendeel waar!

6. Conclusie en discussie

In deze tekst hebben we geprobeerd aan te tonen dat odds en oddsratio's een goede statistische maat zijn om evoluties in een ongelijke participatie te kwantificeren. Natuurlijk betekent dit niet dat percentages fout zijn. Uit percentages alleen kunnen gewoon niet altijd duidelijke en/of eenduidige conclusies getrokken worden over het al dan niet toenemen van de ongelijkheid. Zo zullen uit dezelfde tabel soms tegenstrijdige conclusies getrokken worden. Als er gewerkt wordt met odds en oddsratio's is dit niet meer het geval. De conclusies zullen steeds dezelfde zijn.

In een volgende stap toonden we hoe een logistische regressie ook een odds en oddsratio's als resultaat heeft. Die logistische regressie kan bijgevolg gebruikt worden bij het inschatten van de evolutie van de ongelijke participatie en de significantie ervan. We toonden hoe de interpretatie van de resultaten analoog was aan de berekeningen op basis van de tabel. De informatie over de netto-effecten van meerdere predictoren in het model en hun significantie (inclusief die van de interactietermen) bevestigde de meerwaarde van de logistische regressie.

Deze logistische regressie had nog verder uitgebreid kunnen worden met andere onafhankelijke variabelen, om een duidelijker beeld te krijgen van de netto-effecten. In het voorbeeld dat we doorheen deze tekst gevolgd hebben, kan bijvoorbeeld gedacht worden aan leeftijd. Ouderen (bvb. 60-plussers) hebben minder vaak betaald werk. Het verschil volgens het al dan niet hebben van betaald werk, zal zeker ook ten dele verklaard kunnen worden door de leeftijdsstructuur van de betrokken categorieën. De opname van leeftijd in het model sluit echter niet uit dat we tegelijkertijd nagaan of het verschil volgens het al dan niet hebben van betaald werk nog toeneemt. Zo kan onze analyse nog performanter worden. In dat geval wordt een eigen berekening op basis van de tabel echter moeilijk of zelfs onmogelijk en moeten we direct naar een logistische regressie overstappen. Deze tekst had

echter vooral de bedoeling om odds en oddsratio's ingang te doen vinden en had dus niet zozeer een inhoudelijke focus.

De hier gepresenteerde technieken zijn ruim toepasbaar. In de inleiding werden al de voorbeelden van arbeidsdeelname van mannen en vrouwen en cultuurparticipatie van hoger en lager opgeleiden aangehaald. Maar evengoed kan natuurlijk gedacht worden aan sportdeelname, studievoortgang, milieubewust gedrag… Veel variabelen die interessant zijn voor beleidsmatig onderzoek zijn immers categorisch en/of dichotoom.

Tot slot kan nog opgemerkt worden dat de hier gepresenteerde methode om veranderingen in de tijd te meten, in de Engelstalige vakterminologie gekend is als "Pooling Independent Cross Sections Across Time", het samenvoegen van onafhankelijke steekproeven van verschillende jaren in één dataset. Daar tegenover staat de analyse van paneldata, waarbij een vaste groep mensen opgevolgd en verschillende keren bevraagd wordt. Voor sommige probleemstellingen is dat laatste zeker relevanter en informatiever. Voortbouwend op de toepassing in deze tekst kan bijvoorbeeld gedacht worden aan een onderzoek naar "afhakers", mensen die ooit internet gebruikten, maar het nu niet meer doen. Om dat te onderzoeken zijn paneldata inderdaad beter geschikt. Het voordeel van het samenvoegen van onafhankelijke datasets is dan weer dat er geen speciale statistische modellen vereist zijn. Een gewone (logistische) regressie kan en mag gebruikt worden[9]. Bij paneldata moeten er daarentegen speciale technieken toegepast worden omdat de verschillende observaties gecorreleerd zijn (metingen bij dezelfde personen). Lezers die geïnteresseerd zijn in bijkomende achtergrondinformatie over de analyse van paneldata kunnen zich bijvoorbeeld wenden tot Diggle e.a. (2002) of Verbeke en Molenberghs (2004).

[9] De hier gepresenteerde test voor een toename of afname van de ongelijkheid is eigenlijk een variant van wat in de literatuur gekend is als de Chow-test, een test die nagaat of de relatie tussen de onafhankelijke variabelen en de afhankelijke variabele dezelfde is in twee onderscheiden populaties. Meer informatie over die Chow-test vind je bijvoorbeeld bij McKee-McClendon (p. 281-284).

REFERENTIES

APS (2003) *Kwaliteitszorg Statistisch Productieproces. Aanbevelingen.* Brussel: Ministerie van de Vlaamse Gemeenschap.

Diggle, P., Heagerty, P., Liang, K. & S. Zeger (2002) *Analysis of Longitudinal Data. Second Edition*. Oxford: Oxford University Press.

Hosmer, D. & S. Lemeshow (2000) *Applied Logistic Regression, 2nd Edition*. New York: Wiley.

McClendon, M.J. (2002) *Multiple Regression and Causal Analysis*. Prospect Heights, IL: Waveland Press.

Pampel, F. (2000) *Logistic Regression. A Primer*. Sage University Papers. Series: Quantitative Applications in the Social Sciences, 07-132. Thousand Oakes: Sage.

Verbeke, G. & G. Molenberghs (2001) *Linear Mixed Models for Longitudinal Data*. New York: Springer.

BIJLAGE

Uitvoeren van een chikwadraattest in Excel en een logistische regressie in SPSS

Chikwadraattest bij tabel 1 (zie ook bestand chikwadraat_odds.xls op de SVR-website).

In statistische software, zoals bvb. SPSS kan je een chikwadraattest met 2 of 3 muiskliks opvragen. Maar ook in Excel is die test relatief eenvoudig uit te voeren.

Een chikwadraattest vergelijkt een geobserveerde met een verwachte frequentieverdeling. De verwachte frequentieverdeling is in dit geval die van de hypothese van onafhankelijkheid.

	A	B	C	D	E
2					
3					
4	**Geobserveerde frequentiegegevens**				
5					
6			internetgebruik		
7			ja	nee	
8		2001	495	951	
9		2003	665	765	
10	jaar	2005	885	629	
11					
12					
13	**Verwachte frequentiegegevens bij onafhankelijkheid**				
14					
15			internetgebruik		
16			ja	nee	
17		2001	673,592255	772,407745	
18		2003	666,138952	763,861048	
19	jaar	2005	705,268793	808,731207	
20					
21					
22	**Chi-kwadraat toets die beide verdelingen vergelijkt**				
23					
24	p - waarde		1,3517E-38		
25					

In het stukje van het Excelwerkblad dat hierboven weergegeven wordt, zie je eerst de geobserveerde frequentieverdeling (die overeenkomt met tabel 1).

De verwachte frequentieverdeling werd zelf berekend. Die verwachte frequentieverdeling gaat uit van een onafhankelijkheid van jaar en internetgebruik. Dat wil zeggen dat het relatieve aandeel van de celfrequentie in het rij- en kolomtotaal gelijk moet zijn aan het aandeel van het rij- en kolomtotaal in het algemene totaal.

We verduidelijken dit voor de eerste cel (linksboven). De verwachte frequentie voor die cel is gelijk aan:

$$\frac{495+951}{n} \quad \frac{495+665+885}{n} \; n$$

n is het algemene totaal en gelijk aan 4390 (= 495 + 951 + 665 + 765 + 885 + 629), zodat de formule verder kan uitgewerkt worden:

$$0{,}3294 \times 0{,}4658 \times 4390 = 673{,}59$$

Eens alle verwachte frequenties berekend zijn, kan met één commando de chikwadraattest opgevraagd worden.

Dat commando ziet eruit als volgt:

=CHI.TOETS(C5:D7;C14:D16)

waarbij de celaanduidingen aangeven welke verdelingen vergeleken moeten worden. We vergelijken telkens de cijfers van een rechthoek van drie rijen (5, 6 en 7 versus 14, 15 en 16) in twee kolommen (twee keer C en D). Deze formule moet overigens niet ingetypt worden, maar wordt gegenereerd m.b.v. een extra venster dat opspringt bij het kiezen van de functie CHI.TOETS.

Het resultaat is een zeer klein getal: 0,00…135 waarbij die 1 zich 38 cijfers na de komma bevindt. Dat is de kans om gegeven de verwachte frequenties de geobserveerde frequenties te bekomen. Deze kans is zo klein (kleiner dan 0,05) dat we beslissen dat de verwachte frequenties niet correct zijn en de hypothese van onafhankelijkheid dus ook niet. "Jaar" en "internetgebruik" hangen samen, of ook nog: het internetgebruik is significant toegenomen met de jaren.

LOGISTISCHE REGRESSIES IN SPSS (ZIE OOK BESTAND LOGISTISCHE_REGRESSIES.SAV OP DE SVR-WEBSITE)

De logistische regressies werden uitgevoerd in SPSS.

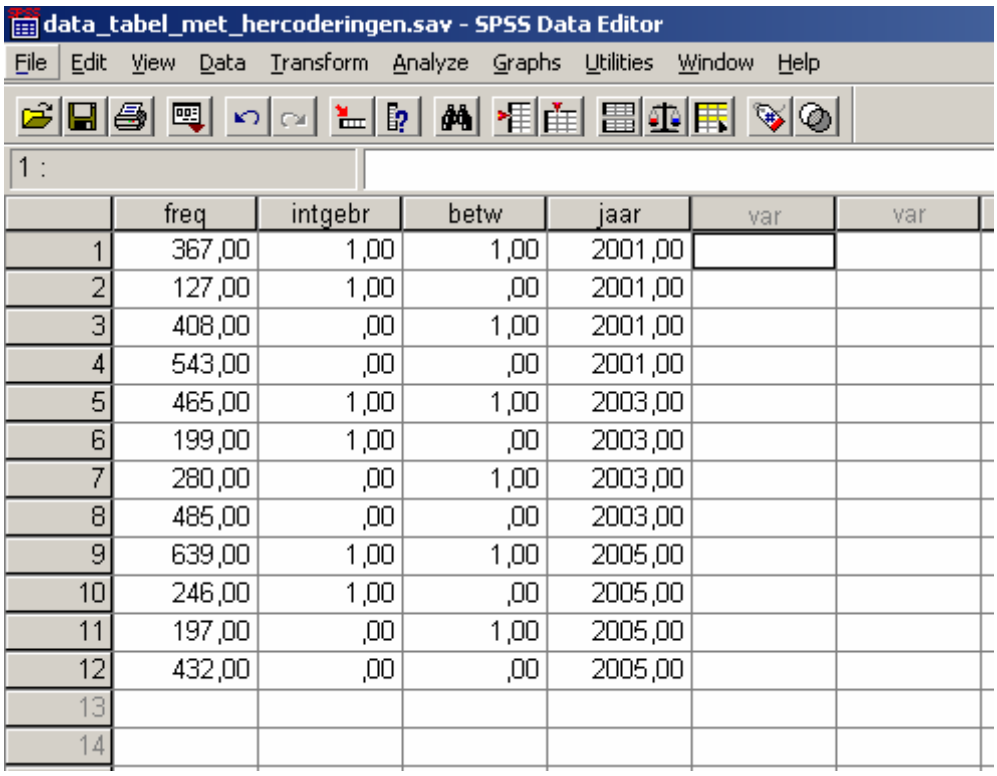
Dat kan voor de atomaire data, of op basis van een tabel (bvb. tabel 2).

Als de ***atomaire data*** beschikbaar zijn, moeten de datasets samengevoegd worden. Dat kan in volgende stappen.

1. Selecteer voor elk afzonderlijk databestand (elk surveyjaar) de variabelen die je nodig hebt
2. Geef die variabelen in elk databestand exact dezelfde naam
3. Voeg in elk databestand een variabele "jaar" toe (die natuurlijk gelijk is aan het jaar waarin de survey plaatsvond)
4. Bewaar de dataset onder een nieuwe naam (bvb. data2001)
5. Voeg de verschillende datasets samen (Merge files – add cases)
6. Voor dit samengevoegde bestand kan je nu dummies aanmaken voor jaar (en zonodig ook nog voor betaald werk en internetgebruik) en ook nog de productterm berekenen via Compute
7. Je bestand is klaar en je kan de logistische regressies uitvoeren

Maar je hebt die atomaire data eigenlijk niet nodig, om te doen wat we in deze tekst getoond hebben. Het kan met behulp van de gegevens van ***tabel*** 2.

1. Eerst voer je die gegevens in in SPSS (Data View) op de wijze zoals hieronder afgebeeld wordt.



	freq	intgebr	betw	jaar	var	var
1	367,00	1,00	1,00	2001,00		
2	127,00	1,00	,00	2001,00		
3	408,00	,00	1,00	2001,00		
4	543,00	,00	,00	2001,00		
5	465,00	1,00	1,00	2003,00		
6	199,00	1,00	,00	2003,00		
7	280,00	,00	1,00	2003,00		
8	485,00	,00	,00	2003,00		
9	639,00	1,00	1,00	2005,00		
10	246,00	1,00	,00	2005,00		
11	197,00	,00	1,00	2005,00		
12	432,00	,00	,00	2005,00		
13						
14						

De eerste kolom bevat de frequenties zoals ze in die tabel staan. Kolommen 2, 3 en 4 geven de waarden weer voor de variabelen internetgebruik en betaald werk en jaar. Internetgebruik en betaald werk zijn al dummy-gecodeerd. De eerste lijn kan je dus lezen als: in 2001 waren er 367 mensen die betaald werk hadden en internet gebruikten. De volgende lijnen spreken dan voor zich.

2. Voor deze dataset kan je ook dummies aanmaken voor jaar ("jaar03" en "jaar05") en producttermen berekenen
3. Nu kan je ook een logistische regressie uitvoeren als je vooraf de data weegt op "freq". De resultaten zullen dan dezelfde zijn als de analyse uitgevoerd op atomaire data.

Het voordeel van de atomaire dataset is wel dat je nog bijkomend kan wegen, bvb. op basis van de gewichten die corrigeren voor non-respons. Het is dan wel noodzakelijk dat je die gewichten mee selecteert in de oorspronkelijke datasets en opneemt in de samengevoegde dataset.

In de reeks SVR – Technisch rapport is reeds verschenen:

- *2006 / 1 Sociaal-culturele verschuivingen in Vlaanderen 2005 Basisdocumentatie*

- *2006 / 2 Bevolkingsprojecties 2004-2025 voor de 308 gemeenten van het Vlaamse Gewest*